Application of a Distribution-Based Assessment of Mission Readiness System for the Evaluation of Personnel Training

David J. Woehr
Michael J Miller
Department of Psychology, Texas A&M University, College Station, TX 77843

and

Winston Bennett, Jr.
U.S. Air Force Armstrong Laboratory Human Resources Directorate
Brooks AFB, TX 78235-5352

ABSTRACT

The present paper summarizes a research project focusing on ways to improve the usefulness of organization level outcome measures of unit readiness/effectiveness. Toward this goal, a measurement approach using unit level outcome measures is presented. The measurement system presented adapts and extends the performance distribution assessment approach proposed by Kane (1986; 1992). We demonstrate that, while originally used with subjective performance judgments, the system is readily adapted to regularly collected unit level outcomes.

A vital concern for the Air Force is the maintenance of mission capability and readiness. A crucial mechanism for the maintenance of mission readiness is personnel training. Of tremendous importance to the design, implementation, and revision of training throughout the Air Force, as with any organization, is the ability to evaluate the effectiveness of training interventions. Specifically, the effective evaluation of any training intervention is crucial to informed decision making regarding the intervention. Central to effective training evaluation is the standard or criteria against which the training is evaluated. In addition, the comprehensive evaluation of training interventions mandates the use of multiple criterion measures.

Organization Level Criterion Measures

Organization level outcome measures represent global indices of effectiveness. While many commonly used criterion measures focus on the assessment of individual effectiveness, organization level measures most often provide more aggregate measures of effectiveness. Recent theories of the criterion construct, have begun to recognize the inextricable relationship between job behaviors and outcomes. Along these lines Binning and Barrett (1989) argue: "... optimal description of the performance domain for a given job requires careful and complete delineation of valued outcomes and the accompanying requisite behaviors" (p.486).

Problems with Outcome-Based Criterion Measures

The detailed delineation of the relationship between job performance and outcomes is especially relevant to training evaluation. An important direction for future research is a focus on behavior/outcome linkages and generating empirical support for these linkages. Unfortunately, the operationalization of specific outcome measures generates somewhat of a dilemma for training evaluation. On the one hand, the ultimate value of training lies in its ability to impact outcomes of value to the organization. Outcome measures (e.g., productivity levels, turnover rates, error rates, etc.) at both individual and aggregate levels would appear to be the ultimate criterion of interest for evaluating training interventions. On the other hand, these measures suffer from a number of problems that limit their usefulness as a standard against which to judge the impact of training.

First and foremost among these problems is the fact that these measures are typically contaminated to an undetermined extent by sources of variance over which the individual has no control. Specifically, the measured outcome is to some extent determined by factors other than individual performance. A second problem with outcome measures is that they are not based on a common metric. Outcome measures are often unique to particular units within an organization and thus are difficult to interpret and compare across organizational work groups or divisions. Additionally, the lack of a common metric typically precludes the meaningful aggregation of performance information across organizational units. A third problem is that these measures only provide an indication of outcome as opposed to the process underlying the outcome. Thus these measures provide little, if any, information about the nature of performance. Finally, the traditional use of outcome measures offers little, if any, means of assessing measurement quality (i.e., how good are the measurements obtained with these measures).

Another major limiting factor with respect to the use of organizational level outcome measures is the lack of conceptual and/or empirical formulations specifying the potential linkages between personnel action and specific outcome measures. For example, if the goal is to evaluate the impact of a particular training program with respect to organizational outcomes it is important to match the nature and content of the training with specific outcome measures likely to be influenced. If the training program focuses on improving maintenance skills then measures most directly related to maintenance outcomes should be identified and examined. Thus while there may be numerous outcome measures available, little if any information exists pertaining to the performance relevance of these measures.

In summary, while organizational level outcome measures are a potentially valuable criterion against which to evaluate training effectiveness, several factors have limited the utility of these measures. These factors include: a) contamination by non-performance related factors; b) lack of a common measurement metric; c) a focus on overall level rather than the performance process; d) lack of any indication of measurement quality; and, e) no conceptual/empirical formulations of the linkage between specific actions and outcomes. Thus, any system that attempts to include outcome measures must address these issues.

Identification of Aircraft Maintenance Related Measures of Performance

One of the primary objectives of the present study was to identify and examine the utility of aircraft maintenance related measures of performance typically collected and used by the Air Force. Measures of performance (MOPs) are qualitative or quantitative measures of system capabilities or characteristics (USAF/TEP, 1994). Toward this end, several sources of data were identified through interviews with supervisory level maintenance personnel. One source of such data was a combination of CAMS- based maintenance data and unit mission characteristic data. This data is routinely collected and reported by aircraft maintenance units as an index of mission effectiveness. This data takes into consideration both equipment and unit mission and manpower characteristics. Example measures include fully mission capable rate (FMC), man hours per flying hours, air and ground abort rates, and delayed discrepancies (deferred maintenance actions). Although a valuable source of information with respect to mission capability, these measures illustrate many of the disadvantages associated with operational measures. More specifically, the metric of each measure is unique to the characteristic being measured. Thus data is difficult to combine and summarize across measures. Further, the measures are cumbersome to summarize. While the measures lend themselves to typical overall summations such as the mean performance level, such measures of central tendency only provide part of the overall picture. Other important information includes the amount of fluctuation and the percent of time at or above some preset standard.

Despite these limitations, these indices have many desirable characteristics with respect to training evaluation. These characteristics include:

1. The measures are regularly and systematically collected.

2. It appears that these indices are both required by and reported to MAJCOM. Thus it is likely that these measures are available Air Force wide.

3. The mission capable/readiness indices reflect both equipment, mission, and manpower characteristics.

4. The indices are easily aggregated from the individual unit level to higher levels of the organization (wing, command, etc.).

5. The indices reflect multiple measures of performance within a specified time span (iterated job function) and thus are readily amenable to the DEAMR system.

Distributional Approach to Criterion Measurement

A second objective of the present study was to develop and evaluate a measurement system that increases the utility of regularly collected operational measures of performance. Toward this end a specific measurement system is presented. The measurement approach presented here extends the system for assessing individual performance developed by Kane (1986) to outcome level criteria measurement. It is believed that this approach may offer a partial solution to the problems associated with outcome measures. The original system presented by Kane (1986), labeled Performance Distribution Assessment (PDA), is based on the distributional measurement model postulated by Kane and Lawler (1979). An important characteristic of this model is a focus on the range of performance observed. Specifically, the model stipulates that not only is the level of performance important, but the fluctuation or variance in performance must also be considered. For example, two individuals may both be appropriately characterized as "average performers"; however, if one is consistently average and the other

alternates between very poor and very good, very different pictures emerge with respect to the individuals' performance. Thus performance measurement must assess the range of performance over time. Specifically, performance is defined in terms of the outcomes of job functions that are carried out on multiple occasions within a specified time span (i.e., iterated job functions). Performance can subsequently be represented in terms of the frequency at which various outcome levels occurred within a given time span.

Another important characteristic of the PDA approach is that it incorporates a relativistic scaling of performance information. More specifically, performance is expressed as a ratio of actual performance (as reflected in the performance distribution generated) to a maximum feasible performance distribution. This maximum feasible distribution reflects the highest level of performance attainable given the constraints under which the work occurs. This scaling process serves to express performance in terms of a relative range of potential performance. Thus, the method allows for quantifiably excluding from consideration in the evaluation of performance the range of performance that is attributable to circumstances beyond the performer's control.

The representation of performance in distributional form along with relativistic scaling has several important advantages. First, it allows for a consideration of performance variability as well as average levels of performance. Thus it allows for an assessment of the consistency of performance and the extent to which negatively valued outcomes are avoided. In this way more information is provided regarding the idiosyncratic nature of individual performance. Second, the relativistic scaling process advocated by the PDA process produces measures of the effectiveness of performance on relativized 0-100% scales with common zero and common upper limits of 100%. Thus any given percentage level remains constant in its meaning regardless of the job, division, or even the organizational level in which it occurs. At the same time, the particular outcome measures used to assess performance may be individualized to meet situational demands and organizational constraints. Specifically, if positions have appreciably different content and extraneous-constraint conditions, measures can be scaled to account for these differences.

The PDA approach was originally advocated as method for enhancing performance ratings. Specifically, it was formulated to incorporate subjective estimates of individual performance outcome frequencies (i.e., supervisory ratings of the frequency at which individuals performed at a particular level). However, its focus on the frequency of particular performance outcomes make it particularly amenable to use with more objective outcome measures. Thus, the application of this methodology to the measurement of organizational outcomes using iterative operational measures appears to be a fruitful avenue for research and may serve to increase the utility of these measures in the training evaluation process.

Adaptation of the PDA Approach for Outcome Level Measures

As noted above, the PDA system appears to be well suited for the measurement and scaling of operational criterion measures. For purposes of illustration, Table 1 presents hypothetical evaluation data for the man hour per fly hour measure presented in PDA format. In Table 1 the performance range represents 5 equidistant steps between the highest possible performance outcome (listed as .30 in the Table) and the lowest performance level (listed as 28.70 in the Table). These values are based on an actual count of man hour per fly hour measure over the course of 1 fiscal. The utility weights represent the utility or value to the organization of performance at each of the 5 levels. These are hypothetical values in this case and would be based on SME estimates. The comparison level values represent a "benchmark" distribution. This "benchmark" distribution may represent either an estimated ideal distribution of performance or the actual performance distribution of a comparison unit (i.e., an earlier time frame or another work unit). The diagram in Table 1 shows the relationship between the 2 performance distributions represented by the actual performance values and the comparison level. Based on this information distributional characteristics for the actual performance values are presented (these characteristics may be expressed in either the utility weight metric or the performance level scale). The total performance effectiveness score represents a quantitative index in a percentage index, that relates the actual distribution to the comparison distribution.

This revised approach to the performance distribution model is labeled here as Distribution-Based Evaluation and Assessment of Mission Readiness (DEAMR). This approach extends the beneficial characteristics of relative distribution based performance assessment to organization level outcome measures.

Table 1: Sample Performance Level Frequencies and Distributional Characteristics for the Man Hour per Fly Hour Measure.

Performance Frequencies						Distribution Characteristics		
Perf. Level	**Perf. Range**	**Utility Weights**	**Perf. Level Freq.**	**Perf. Level %**	**Comparison Level %**		**Utility Wt. Scale**	**Perf. Level Scale**
1	28.70	-100	0	0	0	**Mean =**	38.46	3.77
2	21.60	-50	1	8	1	**SD =**	34.83	0.70
3	14.50	0	2	15	10	**Skewness =**	-1.02	-1.02
4	7.40	50	9	69	80	**Kurtosis =**	1.19	1.19
5	.30	100	1	8	9	**Negative Range Score =**		
		Total Obs. =	13			**Total Perf. Effectiveness**	85.09	

More specifically, characteristics of the DEAMR process include:

1. Performance measurement is relativistic. Outcome measures are scaled relative to maximum possible and minimum acceptable performance levels. Performance distributions are relative to some "benchmark" distribution. Thus, measurement considers the extraneous factors that may influence outcome measures.

2. Performance measurement is based on common metric. All measures are expressed in terms of percentages and thus have minimum and maximum points.

3. Multiple measures of performance are provided; performance is described in terms of mean level, consistency, and negative range avoidance. These multiple measures provide more information about the nature of performance and performance problems.

Another important characteristic of the DEAMR system is that it is easily automated. Relatively little data is required in order to calculate the distributional parameters. This data required includes the highest possible and lowest acceptable performance level, an estimate of the utility weight associated with the lowest acceptable performance level, and the actual frequency of performance outcomes at each of the performance levels. This data may then simply be input into computer or spreadsheet based programs to calculate the distributional characteristics.

References

Binning, J.F., & Barrett, G.V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. Journal of Applied Psychology, 74, 478-494.

Kane, J.S. (1986). Performance distribution assessment. In R. Berk (Ed.), Performance Assessment: Methods and Applications. (pp. 237-273). Baltimore, MD: Johns Hopkins University Press.

Kane, J.S., & Kane, K.F. (1992). The analytic framework: The most promising approach for the advancement of performance appraisal. Human Resource Management Review, 2(1), 37-70.

Kane, J. S., & Lawler, E.E. (1979). Performance appraisal effectiveness: Its assessment and determinants. In B. Staw (Ed.), Research in Organizational Behavior, Vol 1. Greenwich, CT.: JAI Press.

USAF (USAF/TEP). (1994). AFI 99-103, Test and evaluation Process, *25* July 94. Washington, DC.: Author.

Back to Table of Contents

INTERNET DOCUMENT INFORMATION FORM

A . Report Title: Application of a Distribution-Based Assessment of Mission Readiness System for the Evaluation of Personnel Training

B. DATE Report Downloaded From the Internet: 06/04/99

C. Report's Point of Contact: (Name, Organization, Address, Office Symbol, & Ph #): **Navy Advancement Center**
ATTN: Dr. Grover Diel (850) 452-1615
Pensacola, FL

D. Currently Applicable Classification Level: Unclassified

E. Distribution Statement A: Approved for Public Release

F. The foregoing information was compiled and provided by:
DTIC-OCA, Initials: __VM__ **Preparation Date 06/04/99**

The foregoing information should exactly correspond to the Title, Report Number, and the Date on the accompanying report document. If there are mismatches, or other questions, contact the above OCA Representative for resolution.